

# Define.xml Content for SEND Hot Topics

**Rachel Harper**  
**Senior Information Scientist**  
**Regulatory Services**  
**Instem**

# Define.xml in a Submission

## 4.1.4.5 Data Definition Files for SDTM, SEND, and ADaM

*The data definition file describes the metadata of the submitted electronic datasets, and is considered arguably the most important part of the electronic dataset submission for regulatory review.*

FDA Study Data Technical Conformance Guide  
March 2017

# Define.xml content

- Describes a collection of datasets for a study
  - What domains are included, with description of each
  - Description of each variable in each domain
  - Description of entries in those variables
  - Controlled terminology used
  - Algorithms used
  - Any other comments that describe the data

# Define.xml content for SEND

- Some content is from the SEND standard
  - Dataset class and purpose, published controlled terminology
- Some content can be created from datasets
  - List of domains, domain label
  - List of variables, variable label, type, length
- Some content is study-specific
  - Dataset Key
  - Variable Origins, Derivations, Comments
  - Custom terminology lists
- SENDIG includes requirements for define.xml content

# Some SEND Content Challenges

- › Origin Values
- › Codelists
- › Value-Level Metadata
- › Comments
- › Keys
- › Versions

The background is a deep blue with intricate, white, fractal-like patterns that resemble smoke or energy flows. A bright, multi-pointed star-like shape is located in the center-right area, with several lines radiating from it. The overall effect is ethereal and dynamic.

Origin Values

# Origin Fundamentals

- Provides traceability metadata
- Defined at the variable or value level
  - Optional at the variable level if defined at the value level
  - If defined at both levels, must all be consistent
- Variables with origin of Derived require definition of the method
  - Usually can be done within the define file
  - Can reference separate documents

# Current Origin Values for SEND Entries

The SENDIG defines the following controlled terms for specifying Origin:

|                |                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| COLLECTED:     | A designation of "COLLECTED" means that the data were collected via manual or electronic means in a nonclinical study setting. This term should be used for data that were actually observed or recorded by a person or received from an instrument; it should not be used for data that have been calculated, interpreted, or derived from other information.                                     |
| DERIVED:       | "DERIVED" is assigned to values that were not directly collected, but were calculated by an algorithm or reproducible rule, which was dependent upon other data values. The derivation is assumed to be performed by the producer of the dataset. This does not apply to derived lab test results performed directly by laboratories (or by devices).                                              |
| OTHER:         | An origin of "OTHER" is used for values whose origin is known but which were neither collected nor derived. Examples could include values sourced from the protocol or which are defined as part of the Trial Design preparation (see Section 7) or values created as a part of the preparation of the datasets, such as STUDYID, USUBJID, DOMAIN, --SEQ, or --TESTCD.                             |
| NOT AVAILABLE: | "NOT AVAILABLE" means that the origin of the data is not available, such as data received from another laboratory, without the associated annotations or sufficient information provided to determine the actual Origin assignment for the data. Sponsors may specify additional details about the origin that may be helpful to the reviewer in the Comments section of the data definition file. |

# Evolution of Origin Values

- Origin Values defined in SENDIG 3.0 and 3.1
- Origin Values in Define-XML standard
  - Define-XML 1.0 did not include requirements for specific origin entries
  - Define-XML 2.0 embeds a **non-extensible** list of clinical origins
    - Errata published to permit use of origins defined in SENDIG  
<https://wiki.cdisc.org/display/PUB/Define-XML+2.0+Errata>
  - Define-XML 2.1 draft will use published controlled terminology
    - Managed as NCI/CDISC non-extensible controlled terminology
    - Published CT takes precedence over definitions in Define-XML standard

# Evolution of Origin in Define-XML

- Define-XML 1.0 Origin as an attribute of ItemDef element
- Define-XML 2.0 Origin is an element with a Type attribute
  - Changes where origin is defined, but not its use
  - Required for regulatory submissions
- Define-XML 2.1 Origin is an element with Type and Source Attributes
  - Currently source attribute is not required for nonclinical; rationale:
    - Test facilities and test sites are all listed in TS domain
    - Test site data can be identified with --NAM variable
    - All other data originates at the test facility

# def:Origin element in Define-XML 2.1

- The *def:Origin* element includes the *Type* and *Source* attributes.
- The *Type* attribute indicates how the data originated
- The *Source* attribute identifies the party responsible for the data's origin.

```
<!-- Item Definition: Variable Level (STUDYID) -->
<ItemDef OID="IT.STUDYID" Name="STUDYID" DataType="text" Length="7"
  SASFieldName="STUDYID">
  <Description>
    <TranslatedText xml:lang="en">Study Identifier</TranslatedText>
  </Description>
  <def:Origin Type="Protocol" Source="Sponsor"/>
</ItemDef>
```

The background is a solid blue color with intricate, white, fractal-like patterns. These patterns consist of many thin, curved lines that swirl and intersect, creating a sense of movement and complexity. The patterns are most concentrated in the center-right area, where they form a bright, star-like shape. The overall effect is a dynamic and artistic representation of data or code.

**Codelists**

# Codelist Fundamentals

- “Controlled Terminology” in the context of a study refers to the set of all allowable values across all variables that have finite sets of allowable values in the study.
- A “codelist” is a unique subset of the controlled terminology to which one or more variables are subject.
- Beginning with SDTM Version 1.2, the SDTMIG and SENDIG require controlled terminology for many SDTM variables.
- For some variables, sponsor-specific controlled terminology is recommended.

From Define-xml 2.1 standard, Section 4.3

# SENDIG

All entries in the SENDIG “Controlled Terms, Codelist, or Format” are not all define-xml codelists

- ISO 8601 is not a codelist – it is a presentation format
- The presentation format of the variable is not in the data definition file
- Allowed entry for DOMAIN variable could be in a single-value codelist
- Variables with no entry can have a codelist

| Variable Name | Variable Label                    | Type | Controlled Terms, Codelist, or Format |
|---------------|-----------------------------------|------|---------------------------------------|
| STUDYID       | Study Identifier                  | Char |                                       |
| DOMAIN        | Domain Abbreviation               |      | DM                                    |
| USUBJID       | Unique Subject Identifier         | Char |                                       |
| SUBJID        | Subject Identifier for the Study  | Char |                                       |
| RFSTDTC       | Subject Reference Start Date/Time | Char | ISO 8601                              |
| RFENDTC       | Subject Reference End Date/Time   | Char | ISO 8601                              |
| SITEID        | Study Site Identifier             |      |                                       |
| BRTHDTC       | Date/Time of Birth                | Char | ISO 8601                              |
| AGE           | Age                               | Num  |                                       |
| AGETXT        | Age Range                         | Char | number                                |
| AGEU          | Age Unit                          |      | (AGEU)                                |

# Which Variables should have a Codelist?

- Variables associated with CDISC controlled terminology – YES
  - However, is there value repeating a published list?
- Variables without CDISC controlled terminology – MAYBE?
  - Is a full list of terms needed to interpret individual values?
  - Would decodes for the valid values facilitate data interpretation?
  - Is the list of terms allowed during collection useful for review?
  - Is a list of one protocol-defined entry useful?

# Report All Values or Values Used?

- SDTMIG, MSG-SDTM indicate that all permissible values should be included whether they are represented in the submitted data or not
  - Similar text not included in the SENDIG 3.0 or 3.1
- Use full list of terms may be needed in order to understand the context of terms used on a study
- Example:
  - Scores from 1 to 10 are permitted
  - Only scores 1 to 5 were recorded on the study
- Current Uses: Severities, some lab tests
- Later Uses: dermal and ocular scoring, FOBs

# Published Codelist Subsets

- Use when a dataset variable uses a discrete subset of terms from a common codelist
- NY Codelist, different possible uses based on CDISC Notes
  - --BLFL, --DRVFL, --EXCLFL, --FAST, --USCHFL allows Y or Null only
  - --SPCUFL allows N or Null only
  - MADTHREL and MIDTHREL allows Y, N, U or Null only
  - TFDTHREL allows Y, N, U only
- UNIT Codelist
  - Subsets of list based on what use for a test or treatment
- SPEC Codelist
  - Separate gross and micro tissue lists

# Current PhUSE Initiatives

- Clinical and nonclinical teams coordinating
- PhUSE Nonclinical Working Group, project: Demystifying define-xml codelists for nonclinical studies
  - Starting shortly
  - Contact Jennifer Feldmann ([Jennifer.Feldmann@instem.com](mailto:Jennifer.Feldmann@instem.com))
- PhUSE Optimizing Use of Data Standards Working Group project: Best Practices for Metadata Documentation (clinically focused)
  - Clinically focused, work in progress
  - Contact Marcelina Hungria ([mhungria@dicoregroup.com](mailto:mhungria@dicoregroup.com))



# Value-Level Metadata

# What is Value-Level Metadata

- *Defines values in variable for a subset of rows in a dataset*
  - *For example, the OMORRES value for organ weight and each organ ratio cannot be described with the same metadata; each can have a different type, length, origin and derivation.*

## Value Level Metadata - OM[OMORRES]

| Variable | Where                                                          | Type    | Length / Display Format | Controlled Terms or Format | Origin    | Derivation/Comment                               |
|----------|----------------------------------------------------------------|---------|-------------------------|----------------------------|-----------|--------------------------------------------------|
| OMORRES  | <a href="#">OMTESTCD</a> EQ OWBR (Organ to Brain Weight Ratio) | float   | 11.5                    |                            | DERIVED   | Organ Weight / Weight of organ with OMSPEC=BRAIN |
| OMORRES  | <a href="#">OMTESTCD</a> EQ OWBW (Organ to Body Weight Ratio)  | integer | 6                       |                            | DERIVED   | Organ Weight / Terminal Body Weight              |
| OMORRES  | <a href="#">OMTESTCD</a> EQ WEIGHT (Weight)                    | float   | 8.3                     |                            | COLLECTED |                                                  |

# Evolution of Value-Level Metadata

## ➤ Define-xml 1.0

- Value-level metadata is assumed to describe the contents of the --ORRES variable (source: define-xml 2.0, section 4.4, paragraph 3)
- Subset of values to describe can be based on the entry in one variable

## ➤ Define-xml 2.0

- Values in any variable can be described; the variable to describe is part of the value-level definition
- Subset of values to describe can be based on a multi-value “Where” statement

| Variable | Where                                                                                                                                                                                  |
|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| LBORRES  | <a href="#">LBTESTCD</a> EQ ACANT (Acanthocytes)<br>AND <a href="#">LBSPEC</a> EQ WHOLE BLOOD<br>AND <a href="#">LBCAT</a> EQ HEMATOLOGY<br>AND <a href="#">LBMETHOD</a> EQ MICROSCOPY |

# When to Use Value-Level Metadata for SEND

- “It is left to the discretion of the creator when it is useful to provide Value Level Metadata and when it is not. The creator should speak to the consumer if they would like further guidance.”

Define-xml 2.0, Section 4.4

- No information in Study Data Technical Conformance Guide
- Select a variable to define
  - Commonly --ORRES in Findings and --VAL in supplemental qualifiers, TS, TX
- Select variables that define the subset of data
  - Commonly TESTCD for most findings datasets; additional variables for LB and PC
  - In SEND 3.1, subset MI by MIRESMOD to split neoplasm, non-neoplasm codelists
  - Commonly QNAM for supplemental qualifiers, TS, TX



# Use of Comments

# Use of Comments

- Free text to further describe the presented data
- Can be defined at the dataset, variable or value level
- Some descriptions are required by the SENDIG
  - When the IG includes multiple ways represent data, this helps the consumer understand which option was used
- Others may be study or system specific
  - Proactively clarify any ambiguities in data presentation

# SEND Implementation Guide 3.0 References

- SDTM Core Variable: when no data exist for an Expected variable, a null column should still be included in the dataset, and a comment should be included in the **data definition file** to state that data was not collected.
- DM domain, RFSTDTC CDISC Notes: The sponsor must define what collected date is used to populate RFSTDTC in the **data definition file**
- DM domain, RFENDTC CDISC Notes: The sponsor must define what collected date is used to populate RFENDTC in the **data definition file**.
- DM domain, Assumptions: Sponsors should indicate how AGE was populated in the **define file comments**.
- SE domain, Assumptions: If the start date/time of an Element was not collected directly, the method used to infer the Element start date/time should be explained in the Comments column of the **data definition file**.
- EX domain, Assumptions: EXDOSE: The sponsor's **data definition file** should indicate whether the values in EXDOSE represent intended or actual dose levels.

# SEND Implementation Guide 3.0 References

- LB domain, LBNRIND CDISC Notes: Sponsors should specify in the study metadata (Comments column in the **data definition file**) whether LBNRIND refers to the original or standard reference ranges and results.
- LB domain, LBTOX, LBTOXGR CDISC Notes: The sponsor is expected to provide the name of the scale and version used to map the terms, utilizing the **data definition file** external codelist attributes.
- LB domain, Example 5: The following example shows cases of categorical data that cannot be considered as numeric even though in some cases, it appears the data are a number. The allowed values in these ranges should be defined in the **data definition file**.
- PP domain: Information pertaining to all parameters (e.g., number of exponents, model weighting) should be submitted in either the SUPPPP dataset or the **define file**.



Keys

# Good Keys

- Each dataset must have a defined Key
  - Defines uniqueness, Permits single record look-up
- Key is defined as being specific to the study dataset, but...
  - Examples in SENDIG and Define-xml standards include STUDYID in the keys
- Creating a “Good” key for a dataset
  - Don’t include a permissible variable that is not in the dataset
  - Questionable variables for keys:
    - Use of a variable in the dataset with no value in any row
    - Use of a variable that has the same value in all rows
    - A variable that are mirrors another variable in the key (TEST and TESTCD)
- No information in the Study Data Technical Conformance Guide

The background is a solid blue color with intricate, white, fractal-like patterns. These patterns consist of many thin, curved lines that swirl and intersect, creating a sense of movement and depth. The patterns are most concentrated in the center and right side of the image, with some lines extending towards the left edge.

**SEND Study Submissions with Define.xml**

# Define.xml Versions

- Version 1.0 published 5 Feb 2005 and accepted by the FDA until 15 March 2018
- Version 2.0 published 5 March 2013 and accepted by the FDA
- Version 2.1 public review completed, comments currently being addressed

## *Data Standards Catalog:*

| Use                   | Data Exchange Standard | Exchange Format | Standards Development Organization (SDO) | Supported Version | Implementation Guide Version | FDA Center(s)    | Date Support Begins (MM/DD/YYYY) | Date Support Ends (MM/DD/YYYY) | Date Requirement Begins (MM/DD/YYYY) | Date Requirement Ends | Regulatory Reference and Information Sources                                    |
|-----------------------|------------------------|-----------------|------------------------------------------|-------------------|------------------------------|------------------|----------------------------------|--------------------------------|--------------------------------------|-----------------------|---------------------------------------------------------------------------------|
| Study data definition | Define                 | XML             | CDISC                                    | 1.0               | N/A                          | CDER, CBER, CDRH | Ongoing                          | 03/15/2018                     | 12/17/2016 [1]<br>12/17/2017 [2]     |                       | <a href="https://www.cdisc.org/standards/define-xml">CDISC.org - Define-XML</a> |
| Study data definition | Define                 | XML             | CDISC                                    | 2.0               | N/A                          | CDER, CBER, CDRH | 08/07/2013                       |                                | 12/17/2016 [1]<br>12/17/2017 [2]     |                       | <a href="https://www.cdisc.org/standards/define-xml">CDISC.org - Define-XML</a> |

# Considerations for Submission

- Does the file follow the technical standard?
- Does the technical content match the SEND files?
- Are the links to the SEND xpt files correct?
  - No explicit reference to a location
- Is the content required by the standard present and accurate?
- Is the content required by the SENDIG present and accurate?

*Automated tools can be used to check for technical conformance with the standard but cannot check whether all content is accurate.*



Information Solutions For Life

# Define-XML 2.1 Change Summary

- Versioning of Standards and Controlled Terminology
- Identifying non-standard domains
- Identifying non-standard variables
- Enhanced Origin
- Use of Alias for longer SAS names
- XML Schema enumerations
- Fixed Define-XML v2.0 errata
- Other Minor Changes



January 2017 CDISC Webinar recording available:  
[www.cdisc.org/](http://www.cdisc.org/)